Genome Assembly and Annotation of *Pseudomonas aeruginosa* strain PaLo3 Chromosome, Complete genome

SOFIA AMARAL

Introduction

Purpose of the Study

1. To complete a genome assembly and annotation of two sections from a chosen genome.
2. To verify that the organism matches the chosen genome.
3. To identify specific genes in the genome and note their function and location in the organism.

Why This Organism Was Chosen

1. *Pseudomonas aeruginosa* is prevalent in environmental settings like soil and water making it important for ecological studies
2. *Pseudomonas aeruginosa* is highly resistant to antibiotics and an annotation can help understand which genes contribute to its resistance.

Significance

1. Genome assembly provides a starting point to analyze an organism's structure and function.
2. Genome annotation identifies genes and functional elements in a genome, helping understand it's metabolic, virulent, and resistant capabilities.

Methods: Genome Assembly

1. SPAdes V4. 1.0

spades.py -1 SRR33333205_1.fastq.gz -2 SRR33333205_2.fastq.gz -o spadesout

Used to **assemble** genomes from short-read sequencing data. The platform breaks down the sequences into smaller k-mers and used to reconstruct the gene.

2. ABySS v2 3.7

abyss-pe name=assembly k=96 B=2G in='SRR33333205_1.fastq.gz SRR33333205_2.fastq.gz'

Used to **assemble** genomes. The platform works best with larger genome assemblies such as eukaryotic genomes. The platform breaks down the sequence into smaller k-mers, builds a map for the reconstruction, and rebuilds the genome.

3. QUAST V5. 3.0

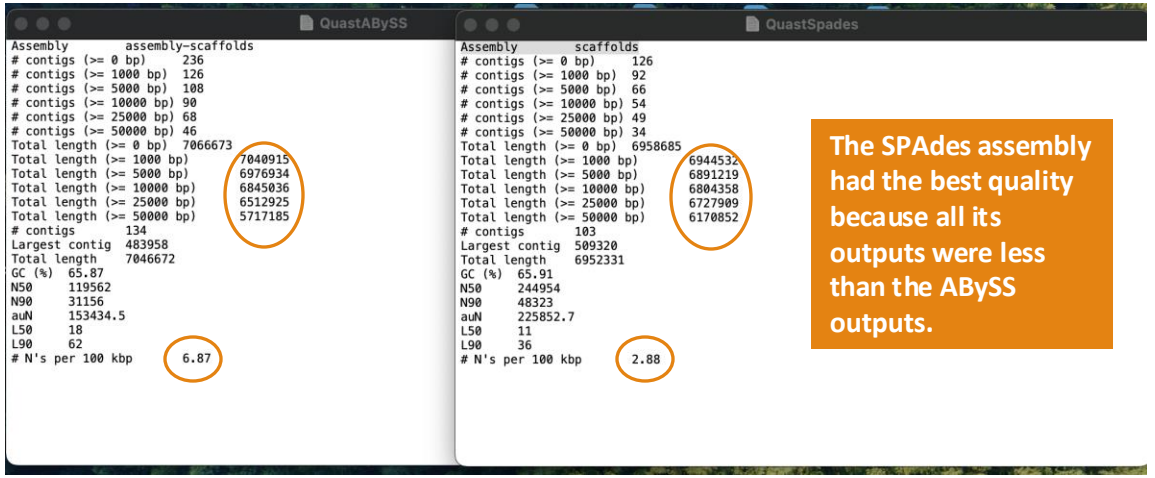
Quast.py spadesout/scaffolds.fasta -o quastspades & quast.py abyss/assembly-scaffolds.fa -oquastabyss

Used after genome assembly to evaluate the **quality** of the assembled genome. In this case, we are comparing the genome assembled through SPAdes and the genome assembled through ABySS.

ABYss OUTPUTS

Name	N	N50	Predicted Genome Length
untigs	509	48	53440
contigs	313	27	101199
scaffolds	236	18	482946

QUALITY COMPARISION



The SPAdes assembly had the best quality because all its outputs were less than the ABySS outputs.

Methods: Genome Identification

1. Barrnap v0.9

barrnap --kingdom bac spadesout/scaffolds.fasta > rRNAsequences.gff

Retrieves the **16S rRNA sequence** so it can later be used to determine the species of the genome. The platform scans the genome for the rRNA sequence looking for specific patterns.

2. Bedtools v2.31.1

bedtools getfasta -fi spadesout/scaffolds.fasta -bed rRNAsequences.gff -fo rRNAsequences.fasta

Used to create a **fasta file** with the 16s rRNA sequences so they can be viewed. The extracted sequence is based on regions specified in the rRNAsequences.gff file created by barrnap.

3. Blastn

Uses the 16s rRNA sequence retrieved from Bedtools to **Identify** the species in the NCBI database.

Methods: Genome Annotation

An annotation involves identifying genes and predicting their **function** in a genome. RAST finds genes in a genome and labels them with their function based on known biology.

- Using the scaffolds.fasta file from the SPAdesout folder several files are generated containing tables with genetic information.
- The parameters for the program were Domain Bacteria, Genus Pseudomonas, Sp aeruginosa, Genetic Code 11, RAST annotation is RASTtk

Methods: FastAni

The platform estimates how **similar** the two neighboring sequences are to the original genome by calculating Average Nucleotide Identity (ANI).

1. Download two related species of *Pseudomonas aeruginosa* strain PaLo3 chromosome, complete genome. The species are Pseudomonas putida and Pseudomonas syringae
2. Create a file called neighbors.txt that contains the two related species.
3. Input the following code to use fastANI for the analysis: fastANI -q spadesout/scaffolds.fasta --rl neighbors.txt -o salmonellaneighbors.txt

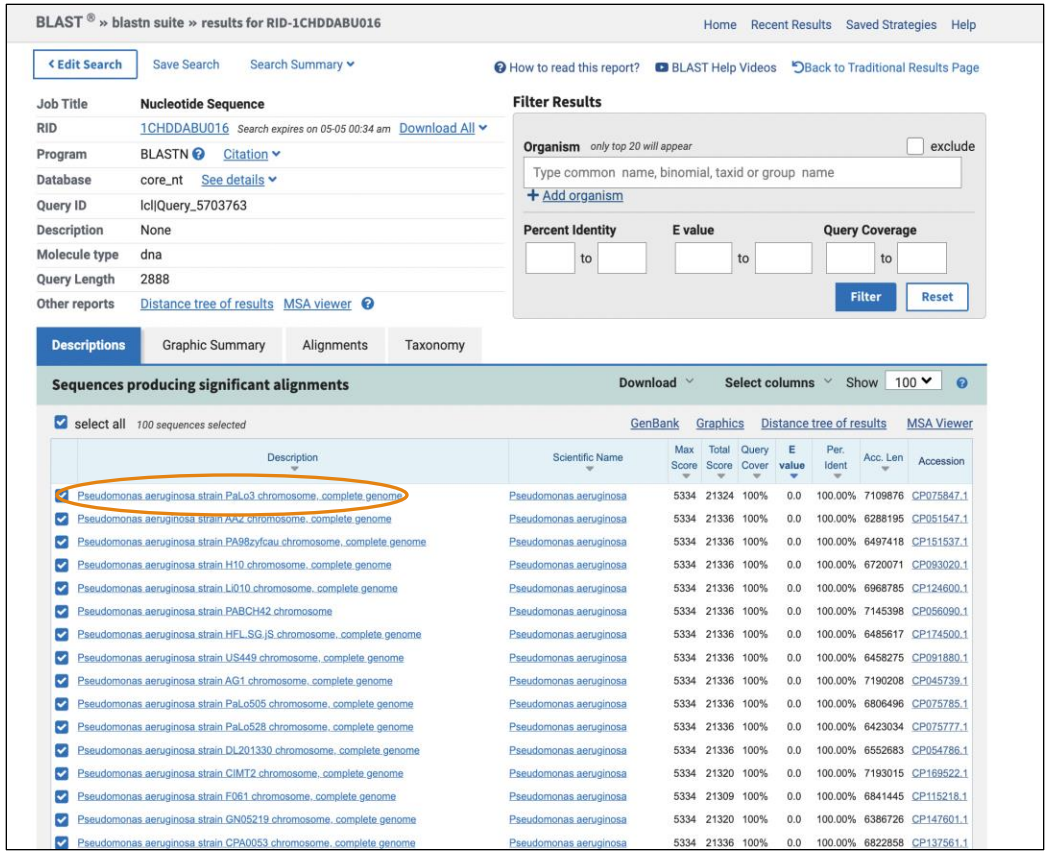
Results: *Pseudomonas aeruginosa* strain PaLo3 chromosome, complete genome

SPECIES IDENTIFICATION

16s rRNA Sequence

>NODE_65_length_5636_cov_87.123797:2455-5343

TCAAGTGAAGAAGCGCATACGGTGGATGCCTTGGCAGTCAGAGGCGATGAAAGACGTGGTAGCCTGCGAAAAGCTTCGGGGAGTCGGCAAACAGACTTTGATCCGGAGA
TCTCTGAATGGGGGAACCCACCTAGGATAACCTAGGTATCTTGTACTGAATCCATAGGTGCAAGAGGCGAACCAGGGGAACTGAAACATCTAAGTACCCTGAGGAAAAGA
AATCAACCGAGATTCCCTTAGTAGTGGCGAGCGAACGGGGATTAGCCCTTAAGCTTCATTGATTTTAGCGGAACGCTCTGGAAAGTGCGGCCATAGTGGGTGATAGCCCCG
TACGCGAAAGGATCTTTGAAGTGAAATCGAGTAGGACGGAGCACGAGAAACTTTGTCTGAACATGGGGGGACCATCCTCCAAGGCTAAATACTACTGACTGACCGATAGT
GAACCAGTACCGTGAGGGAAAGGCGAAAAGAACCCCGGAGAGGGGAGTGAAATAGAACCTGAAACCGTATGCGTACAAGCAGTGGGAGCCTACTTGTTAGGTGACTGC
GTACCTTTTGTATAATGGGTCAGCGACTTATATTCAGTGGCAAGCTTAACCGTATAGGGTAGGCGTAGCGAAAGCGAGTCTTAATAGGGCGTTTAGTCGCTGGGTATAGAC
CCGAAACCGGGCGATCTATCCATGAGCAGGTTGAAGGTTAGGTAACACTGACTGGAGGACCGAACCCACTCCGTTGAAAAGGTAGGGGATGACTTGTGGATCGGAGTG
AAAGGCTAATCAAGCTCGGAGATAGCTGGTTCTCCTCGAAAGCTATTTAGGTAGCGCCTCATGTATCACTCTGGGGGGTAGAGCACTGTTTCGGCTAGGGGGTCATCCCGA
CTTACCAAACCGATGCAAACTCCGAATACCCAGAAGTGCCGAGCATGGGAGACACACGGCGGGTGCTAACGTCCGTCGTGAAAAGGGAAACAACCCAGACCGCCAGCTA
AGGTCCCAAAGTTGTGGTTAAGTGGTAAACGATGTGGGAAGGCTTAGACAGCTAGGAGGTTGGCTTAGAAGCAGCCATCCTTTAAAGAAAGCGTAATAGCTCACTAGTCG
AGTCGGCCTGCGCGGAAGATGTAACGGGGCTCAAACCACACACCGAAGCTGCGGGTGTCACGTAAGTGACGCGGTAGAGGAGCGTTCTGTAAGCCTGTGAAGGTGAGTT
GAGAAGCTTGCTGGAGGTATCAGAAGTGCGAATGCTGACATGAGTAACGACAATGGGTGTGAAAAACACCCACGCCGAAAGACCAAGGGTTCCTGCGCAACGTTAATCG
ACGCAGGGTTAGTCGGTTCCTAAGGCGAGGCTGAAAAGCGTAGTCGATGGGAAACAGGTTAATATTCCTGTACTTCTGGTTACTGCGATGGAGGGACGGAGAAGGCTAGG
CCAGCTTGGCGTTGGTTGTCCAAGTTTAAGGTGGTAGGCTGAAATCTTAGGTAAATCCGGGGTTTCAAGGCCGAGAGCTGATGACGAGTCGTCTTTTAGATGACGAAGTGG
TTGATGCCATGCTTCCAAGAAAAGCTTCTAAGCTTCAGGTAACCAGGAACCGTACCCCAAACCGACACAGGTGGTCGGGTAGAGAATACCAAGGCGCTTGAGAGAACTCG
GGTGAAGGAACTAGGCAAAATGGCACCGTAACTTCGGGAGAAGGTGCGCCGGCTAGGGTGAAGGATTTACTCCGTAAGCTCTGGCTGGTCGAAGATACCAGGCCGCTGC
GACTGTTTATTAAAAACACAGCACTCTGCAAACACGAAAGTGGACGTATAGGGTGTGACGCCTGCCCGGTGCCGGAAGGTTAATTGATGGGGTTAGCGCAAGCGAAGCTC
TTGATCGAAGCCCCGGTAAACGGCGGCCGTAACTATAACGGTCCTAAGGTAGCGAAATTCCTTGTCGGGTAAGTTCCGACCTGCACGAATGGCGTAACGATGGCGGCGCT
GTCTCCACCCGAGACTCAGTGAAATTGAAATCGCTGTGAAGATGCAGTGTATCCGCGGCTAGACGGAAAGACCCCGTGAACCTTTACTGTAGCTTTGCACTGGACTTTGAG
CCTGCTTGTGTAGGATAGGTGGGAGGCTTTGAAGCGTGGACGCCAGTTCGCGTGGAGCCATCCTTGAAATACCACCCTGGCATGCTTGAGGTTCTAACTCTGGTCCGTGAT
CCGGATCGAGGACAGTGTATGGTGGGCAGTTTGACTGGGGCGGTCTCCTCCTAAAGAGTAACGGAGGAGTACGAAGGTGCGCTCAGACCGGTCGGAAATCGGTCGCAGA
GTATAAAGGCAAAAGCGCGCTTGACTGCGAGACAGACACGTCGAGCAGGTACGAAAGTAGGTCTTAGTGATCCGGTGGTTCTGTATGGAAGGGCCATCGCTCAACGGATA
AAAGGTACTCCGGGGATAACAGGCTGATACCGCCCAAGAGTTCATATCGACGGCGGTGTTTGGCACCTCGATGTCGGCTCATCACATCCTGGGGCTGAAGCCGGTCCCAA
GGGTATGGCTGTTCGCCATTTAAAGTGGTACGCGAGCTGGGTTTAGAACGTCGTGAGACAGTTCGGTCCCTATCTGCCGTGGACGTTTGAGATTTGAGAGGGGCTGCTCCT
AGTACGAGAGGACCGGAGTGGACGAACCTCTGGTGTTCCGGTTGTCACGCCAGTGGCATTGCCGGGTAGCTATGTTCGGAAAAGATAACCGCTGAAAGCATCTAAGCGGG
AAACTTGCCTCAAGATGAGATCTCACTGGGAACTTGATTCCCCTGAAGGGCCGTCGAAGACTACGACGTTGATAGGCTGGGTGTGTAAGCGTTGTGAGGCGTTGAGCTAAC
CAGTACTAATTGCCCGTGAGGCTTGACCA



Results

GENOME ANNOTATION

- A genome annotation is important to the study because there is no point in having an assembled genome if you don't know what the function of the genome.
- Genome annotation provide **function** and **location** of a gene helping with data interpretation.
- There are a total of 6958 genes generated from the genome and 3 examples are listed to the right.
  - The examples contain genes that aid in the organism's antibiotic resistance or environmental capabilities.

Feature ID	Fig|6666666.1448083.peg.782	**Fig|6666666.1448084.peg.6257**	Fig|6666666.1448083.peg.455
Location	4404-42547	364518-365492	67098-66223
Length	1458	975	876
Contig	NODE_13_length_186948_cov_28.220173	NODE_1_length_509320_cov_26.828138	NODE_13_length_186948_cov_28.220173
Function	OprM is an outer membrane lipoprotein component of the MexAB-OprM multidrug efflux system.	Polymyxin resistance protein ArnC, glycosyl transferase (EC 2.4.-.-).	Chemotaxis protein methyltransferase CheR (EC 2.1.1.80)

Results

FASTANI

Initial Genome	Reference Genome	ANI (%)	Fragments Aligned	Total Fragments
Spadesout/scaffolds.fasta	Neighbors/putida.fasta	80.2176	901	2269
Spadesout/scaffolds.fasta	Neighbors/syringae.fasta	78.8933	598	2269

Importance to the study: Quality, Accuracy, and Completeness

1.ANI %: A high ANI % supports the identification and quality of the genome assembly by confirming it with bacterium of the same species; insuring the reliability of the annotation.

1.Fragments Aligned: The number of aligned fragments determines the accuracy of the genome. When more fragments are aligned it means the original genome and the neighboring species are related.

1.Total Fragments: Measures the completeness of the genome by comparing the number of aligned fragments relative to the total fragment count. This will determine if the assembled genome has any missing sequences or incomplete annotations.

Conclusion

- The Quast platform proved SPAdes was the better-quality assembly.
- Genome assembly using SPAdes determined the organism was *Pseudomonas aeruginosa* strain PaLo3 chromosome, complete genome.
- Annotations created by RAST listed several genes in the *Pseudomonas aeruginosa* genome along with their function in the organism's resistance and resilience.
- Analysis of the neighboring species compared to the original genome solidified it was a *Pseudomonas aeruginosa* strain.
- Using PathogenFinder v0.4.1 it was predicted *Pseudomonas aeruginosa* strain PaLo3 chromosome as Human Pathogenic.